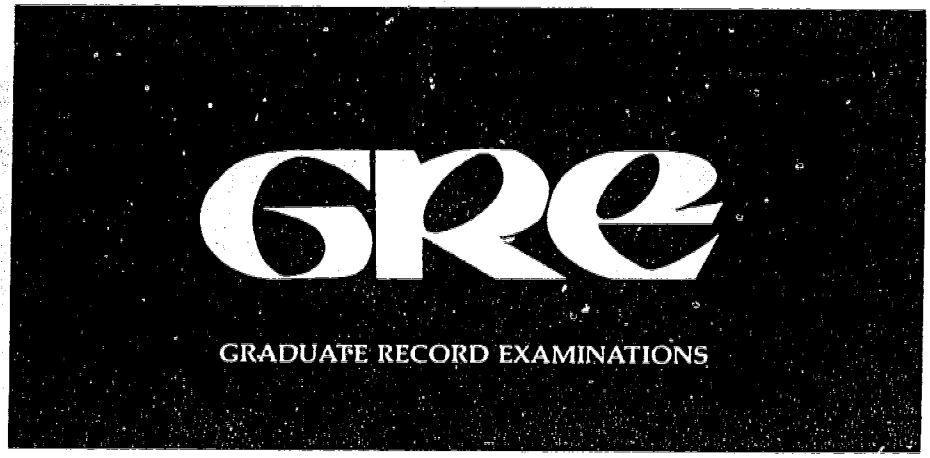
DOCUMENT RESUME

ED 283 829 TM 870 306

| | |
|---|---|
| AUTHOR | Duran, Richard P.; And Others |
| TITLE | GRE Verbal Analogy Items: Examinee Reasoning on Items. |
| INSTITUTION | Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program. |
| SPONS AGENCY | Graduate Record Examinations Board, Princeton, N.J. |
| REPORT NO | ETS-RR-87-5; GREB-PR-82-20P |
| PUB DATE | Feb 87 |
| NOTE | 60p. |
| PUB TYPE | Reports - Research/Technical (143) |
| EDRS PRICE | MF01/PC03 Plus Postage. |
| DESCRIPTORS | *Abstract Reasoning; Cognitive Ability; *Cognitive Processes; College Entrance Examinations; Difficulty Level; Graduate Study; High Achievement; Higher Education; Low Achievement; Problem Solving; *Protocol Analysis; Test Construction; *Test Items; Test Wiseness; *Verbal Tests; Vocabulary |
| IDENTIFIERS | *Analogies; *Graduate Record Examinations |

ABSTRACT

Information about how Graduate Record Examination (GRE) examinees solve verbal analogy problems was obtained in this study through protocol analysis. High- and low-ability subjects who had recently taken the GRE General Test were asked to "think aloud" as they worked through eight analogy items. These items varied factorially on the dimensions of vocabulary difficulty (easy vs. difficult), relationship difficulty (easy vs. difficult), and stem-key correspondence (independent vs. overlapping). A scoring system was developed that included three phases: (1) a cognitive process analysis; (2) a global description of the subjects' strategies; and (3) an evaluation of their performance. It was found that vocabulary and correspondence tended to have multiple effects on how subjects solved analogies and affected whether they solved them correctly. On the other hand, relationship difficulty tended to affect whether or not subjects achieved correct solutions but not how they attained solutions. As expected, high-ability subjects tended to use higher-level or more sophisticated strategies more often than low-ability subjects. However, the use of higher-level strategies was common for both groups of subjects. The results are compared with those reported in the cognitive science literature. In addition, the implication of these results for test development and the usefulness of the protocol methodology are discussed. (Author/JAZ)






# GRE VERBAL ANALOGY ITEMS:

# EXAMINEE REASONING ON ITEMS

Richard P. Duran
Mary K. Enright
Leslie P. Peirce

GRE Board Professional Report No. 82-20P
ETS Research Report 87-5

February 1987

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.




EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

# GRE Verbal Analogy Items:

# Examinee Reasoning on Items

Richard P. Duran

Mary K. Enright

Leslie P. Peirce

GRE Board Professional Report No. 82-20P

February 1987

Acknowledgments

We would like to acknowledge the assistance of the following people in the conduct of the study and in the preparation of this report. Jessie Cryer and Debe Harris provided secretarial support in the various phases of the study. Harriet Johnson provided advice and assistance in planning the recruitment of subjects. Lydia Lesh, Nancy Baker, John Carson, Edward Shea, and Christina Taylor helped in the on-site recruitment of subjects. Barbara Phillips secured score reports and background information on subjects. Finally, Faith Thompson typed the final report, with its many revisions.

Abstract

Information about how GRE examinees solve verbal analogy problems was obtained in this study through protocol analysis. High- and low-ability subjects who had recently taken the GRE General Test were asked to "think aloud" as they worked through eight analogy items. These items varied factorially on the dimensions of vocabulary difficulty (easy vs. difficult), relationship difficulty (easy vs. difficult), and stem-key correspondence (independent vs. overlapping). A scoring system was developed that included three phases: (i) a cognitive process analysis, (ii) a global description of the subjects' strategies, and (iii) an evaluation of their performance. Overall we found that vocabulary and correspondence tended to have multiple effects on how subjects solved analogies and affected whether they solved them correctly. On the other hand, relationship difficulty tended to affect whether or not subjects achieved correct solutions but not how they attained solutions. As expected, high-ability subjects tended to use higher-level or more sophisticated strategies more often than low-ability subjects. However, the use of higher-level strategies was common for both groups of subjects. The results are compared with those reported in the cognitive science literature. In addition, the implication of these results for test development and the usefulness of the protocol methodology are discussed.

Table of Contents

# Introduction

## Overview

The present study represents a collaborative effort by Educational Testing Service test development staff and research staff to improve our understanding of how students taking the Graduate Record Examination (GRE) actually solve one kind of problem on the GRE General Test. The problem domain selected for investigation was verbal analogy items, and the research method of choice was the analysis of think-aloud protocols obtained from subjects as they worked through actual GRE analogy items. It was expected that such an approach would provide information that would be useful to test developers as well as extend our knowledge of the cognitive processes that underlie performance on analogy items.

In the following pages, the nature and purpose of GRE analogy items are described. The potential benefits of this research from the perspective of test development and cognitive science are discussed. Next, relevant research about how analogies are solved is summarized. Finally, the research questions that guided the present inquiry are detailed and the experimental approach described.

## GRE Verbal Analogies

Verbal ability, as measured by the GRE General Test, is defined as the ability to reason with words in solving problems. While knowledge of individual words is an important component of verbal ability as it is conceived for prospective graduate students, verbal reasoning also involves the ability to perceive, analyze, and apply relationships among words or groups of words and longer written passages.

The verbal ability measure in the GRE General Test is composed of four item types: antonyms, analogies, sentence completions, and reading comprehension. The first three of these item types are often referred to as "discrete verbal items" since each item is based on a unique stimulus. As the amount of context provided in the stimulus of each item type increases--from a single word in antonyms to a pair of words in analogies, to a sentence in sentence completions, and finally to a reading passage of 150 or 450 words in reading comprehension--additional skills beyond vocabulary knowledge are required. In particular, the importance of discerning relationships, making inferences, and evaluating and applying information increases.

GRE analogy items would appear to occupy an intermediate position on this continuum. Analogy items are composed of a given a pair of words, the stem, and five options, each consisting of a pair of words. Examinees are instructed to choose the option whose pair of words best expresses a relationship similar to that expressed in the original pair. A GRE analogy item is chosen for inclusion in a test on the basis of its statistical parameters (determined by pretesting) and certain content and psychometric characteristics called for by test specifications.

## The Test Development Perspective

From the perspective of test development staff, greater knowledge of the cognitive skills required in the solution of GRE analogy items would be of immediate as well as extended benefit.

Virtually all verbal items are written by test development staff. Item writers can and do tailor items specifically to meet content specifications, but they can regulate item difficulty only approximately. Among the analogy item features that an item writer can most easily manipulate to influence item difficulty are the level of vocabulary difficulty of the individual words, the complexity of the relationship between the given words, and the degree of difficulty in eliminating the distracters (the "closeness" of the distracters). Particularly with regard to the latter two features, an item writer relies primarily on intuitive judgment as to how complex examinees are likely to find a particular relationship or how "close" they are likely to find a particular distracter. This judgment is based largely on the item writer's having observed the statistical performance of past items and inferring what might have contributed to the observed item parameters. Knowledge of how examinees approach the solution of items would show to what degree test development assumptions about examinee performance are correct.

By providing evidence of how examinees actually go about solving analogy items and what aspects of items appear to contribute to item difficulty, research can help suggest ways in which test development staff could more efficiently produce items with desired content and psychometric characteristics. As with many item pools, it is not easy to develop analogy items in the upper range of the difficulty scale that discriminate satisfactorily. Item writers try, insofar as possible, to develop high-difficulty items that do not depend excessively on the use of arcane words since it is felt that, although GRE analogy items are verbal reasoning items and do depend extensively on word knowledge, antonym items are a more appropriate place to test vocabulary knowledge per se. For example, information about the influence of vocabulary on item difficulty would be helpful to item writers, as would information concerning the interaction of vocabulary difficulty and relationship difficulty. In addition, evidence of how and why examinees choose wrong answers would help item writers in the development of effective distracters, a prime means of influencing item difficulty.

## Cognitive Science Perspective

Recent developments in many areas of cognitive science research indicate that it is possible not only to understand how humans organize information processes in performing cognitive tasks, but also to understand how individual differences in general cognitive aptitudes are related to information processing differences (Glaser & Pellegrino, 1978; Hunt, 1978;

Sternberg, 1982; Whitely, 1980). The quest for knowledge of the latter sort is just beginning and it holds promises for the development of cognitive foundations for the sorts of general aptitudes measured on testing instruments such as the GRE General Test and, in particular, on the verbal ability section of the test.

The pursuit of this quest for us is best founded on the study of information processing strategies underlying a single variety of item type found on the verbal ability subtest. By focusing on a single item type such as analogies, we can draw on specific cognitive research that describes in detail the kinds of problem solving required by this item type. In addition, individual differences in strategies that contribute to success in solving this item type can be observed.

Because of the exploratory nature of the present study and the need to focus in detail on specific processing issues, this study does not examine connections between psychometric indices of item difficulty and discrimination, and the strategies utilized by examinees, although attention is given specifically to examinees' ability to solve analogies as they occur on the GRE test. Psychometric research at the item level would be a valuable complement to the present study, but such work would require specialized research designs.

## Research on Solving Verbal Analogies

Psychologists have long been interested in how analogies are solved, but recent research has been stimulated by the rise of cognitive psychology in the late 1960s and the development of cognitive components theory in the 1970s (Sternberg, 1977). Connolly and Wantman (1964) conducted an important early study analyzing performance on SAT verbal analogy items. Their study is notable because the format of SAT analogies is identical to that of GRE analogies. Connolly and Wantman employed protocol analysis techniques to categorize nine subjects' problem-solving performance on 25 analogy items. Sixty-seven different performance categories were utilized. These categories intermixed evaluations of performance with descriptions of strategies used by subjects. Sample categories included "skips unfamiliar terms," "justifies choice with incorrect logic," "reads stem and immediately states keyed response," and "appears to be guessing between two possibilities." Of the 67 categories, only 14 were used frequently in protocol coding.

Connolly and Wantman found a propensity among some subjects to pay special attention to the first answer option in each analogy problem. Subjects showing this propensity often revised the relationship they had initially inferred for the stem pair so it fit the characteristics of the first option pair. They also found that some subjects selected the key (i.e., correct pair of words) although they showed a clear misunderstanding or an imprecise understanding of the meanings of words in the keyed

to follow a simple generate-and-test problem-solving approach in which they consecutively evaluated the fit of options given the target relationship required. In contrast, on hard items, subjects were more prone to evaluate the options before arriving at the stem relationship necessary for solving a problem. In describing this work, Pellegrino and Glaser (1980) state:

> The protocol data revealed that a process of successively refining the rule consistently occurred across the alternative set, and the extent to which this process was involved was a function of the degree of precision in originally defining the rule and the potential answer. (p. 210)

Heller and Pellegrino found some evidence that appeared to distinguish strategies utilized by the best and worst problem solvers . . . Highly skilled persons tended to articulate more clearly reasons why incorrect options were rejected. Another finding of the Heller and Pellegrino work was that detection of semantic appropriateness of answer options given the meanings of terms making up the stem of verbal analogy problems was significantly related to success in selecting correct responses. The more stem and option words were related to each other in their meanings and associations, the more likely subjects were to pick the correct answer option. This result also supports the earlier findings of Rumelhart and Abrahamson (1973), who found that subjects were sensitive to the semantic relatedness of terms in their interpretation of the appropriateness of verbal analogies.

Whitely and Barnes (1979) investigated whether subjects' requests for information in an analogy simulation task matched various cognitive component models of verbal analogies proposed by Sternberg (1977). Sternberg had postulated that six cognitive components were implicated in analogies solution: encoding, inference, mapping, application, justification, and preparation and response. Encoding components were implicated in the recognition of the meaning of words in an analogy problem. Inference, mapping, and application components were involved in inferring relationships among words, tailoring a relationship between two words so that they might extend to a third term, and extension of a given relationship between two words to see whether it applied to another pair of words. Justification components were involved in evaluating which answer options best completed analogical relationships derived for the initial A:B component of analogies. Finally, preparation and response components were entailed in the production of an answer response to a problem.

Sternberg (1977) in his early work had hypothesized that, although the ordering of these components was invariant, the manner in which components were executed might vary. If a component process was carried out in an exhaustive mode, it would involve thorough consideration of all problem elements that were available and appropriate to execute that component once and for all for the entire problem. In contrast, in a self-terminating mode, a component process would cease as soon as disconfirming evidence

emerged that it would not work for some element of information under consideration.

Whitely and Barnes found that their data supported the conclusion that subjects tended to ask for information in the order postulated by Sternberg and that they utilized a good deal of self-terminating execution of components. In particular, they found that some subjects repeatedly asked the experimenter for more and new information about a relationship between the first two terms of an analogy as they considered each answer option in turn. This form of processing could occur only if subjects were executing components in a self-terminating mode rather than in an exhaustive mode.

Whitely and Barnes also noted that subjects followed some strategies that were not predicted by Sternberg in his early theory. They noted, for example, that confirmation that an answer option was correct on occasion seemed to be different from a notion of justification that could only involve deciding which among several answer options was the best.

Whitely and Barnes concluded that their data indicated that there were considerable individual differences in how subjects organized and executed components and that a single general component model of the verbal analogies solution process was an oversimplification. They called attention to differences among individuals in their general approach to problems and to the implications of these differences for the organization and execution of components in analogy problem solving.

There are two general findings in cognitive components research on analogy problems that are worth citing briefly in the context of the present study. One finding by Sternberg (1977) was that highly skilled problem solvers seem to spend proportionally more time on encoding the words occuring in analogy problems than in carrying out inference, mapping, application, and justification components. Skilled subjects also spend less time proportionately in executing responses to analogy problems. Alderton, Goldman, and Pellegrino (1985) found evidence that skilled analogy problem solvers use execution, evaluation, inference, and mapping components relatively more and exhibit a greater propensity to achieve correct analogy problem solutions as a result. Those who are less skilled in solving analogies tend to become distracted more easily in carrying out inference and mapping processes. This finding was replicated by Whitely (1980), who also reported that highly skilled subjects are more efficient at encoding.

Two recent studies conducted by the GRE Program are also relevant to the current project. Wilson (1985) investigated relationships between GRE General Test subscores and undergraduate senior grade point average. He found that performance on analogy and antonym items was most closely associated with GPA for English, history, sociology, political science, education, and economics majors. Performance on analogy and antonym items was correlated less with GPA for chemistry, computer science,

mathematics, electrical engineering, biology, and agriculture majors. In another finding, Wilson also concluded that combined performance on sentence completion and reading comprehension items was a better predictor of GPA than combined performance on analogy and antonym items. Wilson's findings are interesting in that they suggest that analogy and antonym items may require cognitive and academic skills that are more likely to be cultivated in humanities and social science majors than in scientific and technical majors.

Other GRE research, by Ward (1982), suggests that performance on analogy items in a multiple-choice response format is different from performance on a free-response version of these items. Ward found that availability of answer options was a critical factor in examinees' ability to solve items. Given an analogy of the form A:B as X:Y (read: "A is to B as X is to Y"), selection of the appropriate pair X:Y from among answer options depends critically on seeing which answer option pair best preserves the full extent of relationships present in the given pair A:B. When answer options are not given, examinees may construe the relationship in unexpected ways.

## Research Questions and Approaches

Pilot Nature of the Study. The review of the research literature suggests that variations in the problem-solving activities of examinees working GRE analogy problems might be related both to the examinees' abilities and to the characteristics of the items. Protocol analysis and cognitive components research as well as psychometrically oriented research might all contribute to an analysis of these differences. Given that no previous attempts had been made to apply protocol analysis and other cognitive research methods and findings to the study of GRE analogy items, our efforts in the present project were considered exploratory. Discussion of the design of our study begins with a consideration of the general questions for research that emerged from a test development perspective.

Item Dimensions. From the test development perspective, our work was guided in part by an interest in the relationship between item characteristics and examinee strategies. We decided that, rather than focus on item characteristics as they are laid out in the test specifications, we would investigate characteristics of analogy items that seemed more likely to govern the cognitive processes underlying their solutions. Two of the three item characteristics chosen for this study were level of vocabulary and complexity of the stem relationship. These two characteristics were chosen because they are major elements considered by item writers as they develop analogy items and because they are the features of analogy items on which the item-type description and "tips" sections of the Information Bulletin focus. The third item characteristic chosen for investigation was drawn from the specifications and concerned the nature of the association or correspondence between stem and key words (independent or overlapping). This characteristic was deemed relevant to the

study because, among the specifications categories, it appeared to relate most to a comparison of stem and key and thus potentially to the solution of an item.

Each of these three item characteristics (vocabulary level, complexity of key-stem relationship, and stem-key correspondence) was treated dichotomously, thus yielding eight possible combinations of item characteristics. The two dimensions for each variable were as follows: easy vs. difficult vocabulary, easy vs. difficult relationship, independent vs. overlapping correspondence.

Cognitive Analysis of Protocols. Two distinct but complementary approaches to cognitive analysis of analogy problem solving guided our work: protocol analysis of problem-solving behavior and cognitive component research. A great deal of previous problem-solving research has used think-aloud protocols generated by subjects as they work problems as the source of data for analyses of problem-solving strategies (see Newell & Simon, 1972; and Ericson & Simon, 1984, for a review of this area). Analyses of protocols have most often stressed identification of subjects' problem-solving plans, including establishment of goals in a plan hierarchy and implementation of strategies to evaluate and achieve progress towards goals. Proper use of protocol analysis methods requires application of a theoretical model of problem-solving behavior that can be used to interpret the mental acts referred to by utterances in a protocol. In this study, it was postulated that some of the insights of current cognitive process models of analogy problem solving would be helpful in identifying some of the problem-solving strategies alluded to by subjects who were asked to think aloud as they worked GRE analogy items. Use of a process approach to guide protocol analysis in this way was exploratory.

Cognitive component approaches to verbal analogy problem solving investigate the way in which subjects organize and execute important problem-solving activities. Many research studies in this area use subtractive factor experimental design techniques involving reaction time measures of performance under various task conditions to demonstrate the existence and organization of processes. Our approach differed in that we sought to find evidence for the occurrence of processes in the protocol utterances of persons working GRE verbal analogy items. We sought evidence for the occurrence of various cognitive processes similar to those described in cognitive components research that would be congruent with other evaluations we made of subjects' strategies and problem-solving effectiveness.

Our review of the research literature on analogy problem solving led us to postulate five classes of cognitive processes (Pellegrino & Glaser, 1982; Sternberg, 1982). Encoding processes are enacted when persons read and understand the meaning of individual words in analogy problems. Inference and comparison processes center on inference of relationships among the words making up the stem and options of analogy problems. The relationships inferred might be between terms that are explicitly paired or between words that are not explicitly

paired in the statement of a problem. Inference and comparison processses are at the heart of verbal analogy problem solving. To solve GRE analogy problems, examinees must identify the relationship between the stem word pair, and they must infer relationships that occur among the word pairs making up the options. In inferring relationships, persons may also engage in identifying relationships between words that are not explicitly paired in a stem or option pair. For example, the research literature suggests that after they identify a relationship that holds for the stem pair of words A:B, many subjects go on to infer a relationship between the first word of the stem pair and the first word of the pair of words that would complete the analogy. In the cognitive component research this skill is termed "mapping." More generally, subjects might apply and test a relationship between one word and any other word in an analogy problem.

Decision and response selection processes underlie the critical judgments and choices that examinees must make in narrowing down the possible answers to an analogy problem. To solve GRE analogy problems, examinees must determine which option word pair best exemplifies the appropriate relationship between words in the stem pair. Experimental studies of cognitive processes in verbal analogy problem solving have rarely given attention to this matter in a manner that captures the complexity of problem-solving required in solving analogy items on standardized multiple-choice tests (Goldman & Pellegrino, 1984). Thus, in the context of our study, we sought to create a range of decision and response selection subcategories used by examinees in a multiple-choice testing context.

Confirmation processes in this study allude to examinees' statements of rationales for the selection of answer options. In cognitive process research this component is sometimes treated as an evaluative process, that is, in effect, a final mental check on a problem-solving decision. We wished to be sensitive to variations in the form and content of confirmatory comments. In addition, the very act of requiring examinees to speak about their problem solving might influence some to confirm their decisions more than others.

We saw a need to specify one further process that has not been researched extensively in cognitive process research on verbal analogies--executive processes. Executive processes (cf. metacomponents, Sternberg, 1982) refer to the general strategies that guide and evaluate the effectiveness of an approach to problem solving. Execution of such components would appear to be critical to the solution of GRE verbal analogy problems since examinees must determine the general kinds of criteria that underlie their motivation, allocation of attention, planning of detail, and monitoring of progress while working test items. The need for executive processes is particularly apparent in the case of solution of difficult analogy items, when examinees' must decide on answer options under conditions of uncertainty and ambiguity.

The value of an expansion of process descriptions to describe solution of analogy problems that occur on the GRE is consistent with

suggestions made by Goldman and Pellegrino (1984). In a summary of two previous studies employing protocol analysis to study verbal analogy problem solving, they stated (p. 163)

> ...Difficult items often demand a refinement or redefinition of individual terms and relationships in order to achieve a solution. This involves various amounts of recursive processing. The possibility of such events is generally handled by the concept of a justification component. However, this general component can subsume a number of possible recursive processing sequences. Solution of items representing the difficulty levels found on actual standardized tests may involve significant amounts of such recursive processing. How such a complex process is monitored and its interaction with individual subject characteristics remain largely unexplored areas.... The limited eye-movement data that are available (Bethell-Fox, Lohman and Snow, 1982) suggest that it is inappropriate to conclude that all analogy solution conforms to simple linear process sequences that have been verified chronometrically for proficient adult reasoners.

Global Descriptions of Examinee Performance. In addition to a cognitive process description of examinee solutions of analogy items, we were interested in more global descriptions of examinee performance. Such descriptions were considered to be of value in supplementing the molecular analysis provided by the cognitive process description and of more immediate relevance to test development concerns.

The first area of interest was an overall description of an examinee's approach to the solution of an item. It seemed reasonable to assume that there would be differences among examinees in the overall strategy and sequence of steps in item solution and, possibly, differences in overall strategy and sequence of steps among different classifications of items. For example, it was thought that some examinees would first state the relationship between the stem words, apply that relationship to some or all of the five options, and then choose an answer, and that other examinees would follow a different strategy.

In addition, we were interested in more general descriptions of the extent to which examinees engaged in certain of the cognitive processes described above. Questions we regarded as relevant here included whether or not examinees applied an overall criterion (for example, match of stem and option relationships) to the solution of an item; whether or not they attempted to justify their final solution; how they dealt with an item in which they found two or more options to be equally plausible; and how they dealt with an item in which they did not know the meaning of some words or could not state the relationship between the stem words.

Finally, we were interested in an evaluation of examinees' performance with respect to vocabulary knowledge and ability to define relationships,

and the relationship of these factors to correct or incorrect solutions of items. We hoped to learn what led to correct solutions--was it most often a combination of strong vocabulary and a properly stated relationship, or was one of these factors more important than the other? Equally relevant was the issue of what led to an incorrect solution. Knowledge of the roles played in correct or incorrect item solution by vocabulary knowledge and ability to formulate word relationships and their interaction was considered to be of direct practical benefit to item writers; hence our decision to add this final evaluative analysis of examinee performance to the descriptive analyses outlined above.

## Research Plan

The research plan for this study evolved from the collaboration of test development and research staff and incorporated the concerns and interests of both. While the methodology was drawn from cognitive component research, the variables studied and the specific issues addressed were those that were thought to be of particular interest to test development.

Protocols were obtained from subjects who were asked to solve a set of analogy items aloud. These subjects had recently taken the GRE General Test. The analogies they were asked to solve were drawn from the pool of disclosed GRE items and varied on the dimensions of vocabulary difficulty, relationship difficulty, and stem-key correspondence. The protocols were analyzed on a molecular level to describe the processes subjects used in detail and on a more global level to describe their overall strategies and the quality of their solutions.

## Method

### Experimental Design

The primary questions of interest were how ability level and problem characteristics affect the way examinies solve analogies. First, subjects were tested on a paper-and-pencil analogy test to determine their ability to solve analogies in the format that is used on the GRE test. Then subjects solved eight additional analogies out loud. The analogies for this part of the experiment consisted of eight types of analogies that represented the factorial combination of three bidimensional stimulus characteristics. Three independent groups of subjects, composed of equal numbers of high- and low-ability subjects, were tested on different sets of exemplars of the eight analogy types, thus producing three replications of the experimental design.

### Subjects

Thirty-six college students or graduates who had taken the GRE General Test in December 1983, April 1984, or June 1984 were assigned to the three experimental groups so that the mean GRE verbal scores for the groups were approximately equivalent. There were five males and seven females in two of the groups and four males and eight females in the third. Two additional subjects were tested, but their data were lost due to error in the experimental procedure.

Subjects were recruited by two methods. Initially, flyers describing the experiment were distributed to people leaving GRE test administrations at Rutgers and Princeton universities in December and in April. In addition, letters soliciting participation were sent to individuals who had taken the GRE in June and who resided within a reasonable distance of the experiment site (Educational Testing Service in Princeton, NJ).

GRE verbal scores for the subjects participating in the experiment ranged from 310 to 800, and the mean for all 36 subjects was 540.83 (SD = 134.85). The mean GRE verbal score for recent test forms is approximately 480. This difference between our sample and the overall population reflects the fact that relatively few low-scoring subjects volunteered to participate. Subjects were paid $30 for their participation in the experiment and were reimbursed for any travel expenses incurred.

### Apparatus and Materials

A paper-and-pencil analogy test was administered. In addition, a series of analogy items typed on individual index cards were given to the subjects. A cassette tape recorder was also used.

Stimuli

The paper-and-pencil analogy test consisted of 22 items from the pool of disclosed GRE analogies. The test was administered to all subjects, who were given 11 minutes to complete it. The test was constructed to be parallel in content specifications to the analogy items included in the verbal sections of an operational GRE General Test (the verbal score is based on two test sections totaling 76 items, of which 18 are analogy items). The paper-and-pencil test was constructed to be slightly harder than the analogy sections of recent test forms because of the higher GRE verbal score level of the 36 subjects; the test was made up of 22 rather than 18 items in order to include some additional harder items.

For the second experimental task, three sets of ten analogies were constructed. The first two items were the same for all three sets and functioned as warm-up trials. The next eight items were different for each set. These items varied factorially on three dichotomous stimulus dimensions: (i) vocabulary (easy vs. difficult), (ii) relationship (easy vs. difficult), and (iii) stem-key correspondence (overlapping vs. independent). Two of the experimenters independently classified the items in the pool of disclosed GRE analogies on these three stimulus dimensions. Experimental items were selected from those on which there was agreement between the two experimenters. Different exemplars of the eight possible combinations of these dimensions were selected for each stimulus set, producing three parallel sets of items. The three sets of items were selected to be as equivalent in difficulty level as possible. The three sets of items were equivalent in difficulty and r-biserial (item-test correlation). The mean difficulty of each set fell between delta 12.0-12.1 (delta is a standard deviate with a mean of 13 and a standard deviation of 4), and items in each set ranged from delta 8-9 to delta 15. The mean r-biserial of each set fell between .50-.51 (the mean delta and r-biseral for the GRE verbal sections are 11.9-12.1 and .50-.54, respectively).

A copy of the paper-and-pencil test and of the three sets of analogies can be found in Appendix A.

Procedure

The experiment consisted of two tasks. First, subjects were allowed 11 minutes to complete a 22-item analogy test. For the second task, the subjects solved a set of 10 analogies out loud. This set consisted of two practice items and eight experimental items. The experimental items were administered in a different random order to each subject.

This second task was divided into four phases. First, subjects were told that they were to think aloud as they worked through each problem.

Subjects were instructed to say anything that came to mind even if it seemed irrelevant. However, it was pointed out that the experimenter was particularly interested in information about the subjects' ideas concerning the relationship between words in a pair, their knowledge of the vocabulary, and how they were deciding what the correct answer was. While subjects were solving the analogies, the experimenter confined her comments to reminders to think aloud. Subjects were not informed during this phase whether or not their answers were correct, and there was no time limit for solving the items.

In the second phase of the task, subjects were asked to arrange the eight experimental items in order of perceived difficulty and to describe why some items seemed more difficult than others. Next, some subjects were asked to review the analogies and to explain more fully why they had eliminated some options. This request was made only when the experimenter thought a subject had not been very thorough in the first phase. Finally, all subjects were asked if they wished to know the correct answer and rationale for any item. The entire test session took between one and two hours. Subsequently, typists prepared verbatim transcripts of the experimental session.

## Protocol Coding

Only those sections of the transcript that involved the initial solution of the analogy items in phase 1 were coded. The subjects' solutions for the various items were randomly assigned to the three experimenters for coding. The assignment of items was randomized across subject groups and items so that each experimenter coded material from almost all the subjects and some examples of each item.

The coding was composed of three distinct phases that represented alternative ways of describing a subject's performance: process analysis, global description of strategies, and evaluation of performance.

The process analysis represented a molecular analysis of the protocol, phrase by phrase, and was based on processes identified in previous research (Pellegrino & Glaser, 1982; Sternberg, 1982). The behaviors classified were the explicit statements made by the subjects. There were five scoring categories in the process analysis:

1. encoding comments--statements or phrase that concerned the meaning of words

2. inference and comparison comments--statements that included a subject's descriptions of or inferences about the relationships between words and the application of these inferred relationships to other sets of words

3. decision and response selection comments--statements concerning the likelihood that an option was correct or the selection or elimination of options as the answer

4. confirmation comments--explicit rationales given following the selection of an answer

5. executive comments--statements concerned with planning and carrying out a chosen course of action

The second stage of coding involved relatively more global judgments of the subjects' strategies. In this stage, consideration was given to a subject's overall strategy in solving an item, to whether or not the subject applied a criterion in achieving the solution, to whether or not the subject justified his or her solution, and to the strategies the subject employed (or did not employ) when there appeared to be more than one plausible answer to an item and he or she had difficulty understanding some words or the relationship between stem and key.

In the final stage of coding, each subject's performance was evaluated in terms of his or her knowledge of the meaning of the words in an item and the correctness of his or her delineation of the relationship between stem and key. In addition, the subject's performance was evaluated in terms of what led to a correct or an incorrect solution of an item. A detailed description of the coding system is presented in Appendix B.

A subset of six analogies was coded by all three experimenters to assess intercoder reliability prior to the actual coding. The reliability data for the three sections of the coding system are presented in Table 1. For the process analysis, intercoder correlations were computed for the frequency of occurrence of each type of statement provided that at least one of the three experimenters had coded that type of statement as having occurred at least once. Thus, the correlations in Table 1 are a conservative estimate of interscorer reliability in that data for the coding categories that all three coders agreed had not occurred was not included in the reliability estimate. As can be seen in Table 1, the correlations among the coder pairs for 126 scoring items ranged from .73 to .78. For the "Description of Strategy" and "Evaluation" sections, the percent agreement for the applicable categories coding ranges from 63 percent to 95 percent.

## Results

### Overview

First, the comparability of the three groups of subjects in terms of their GRE verbal scores, their scores on the paper-and-pencil analogy test, and their performance on the experimental analogies is discussed.

Table 1

Intercoder Reliability for Coding System

Process Analysis

Correlations between Coders

| | | Coder 1 | Coder 2 |
|---|---|---|---|
| Coder | 2 | .74 | |
| | 3 | .73 | .78 |

Description of Strategy

Percent Agreement between Coders

| | | Coders 1 | Coders 2 |
|---|---|---|---|
| Coder | 2 | 95% | |
| | 3 | 78% | 75% |

Evaluation of Performance

Percent Agreement between Coders

| | | Coder 1 | Coder 2 |
|---|---|---|---|
| Coder | 2 | 78% | |
| | 3 | 63% | 70% |

The major thrust of the data analysis, however, was to determine how analogy solution was affected by subject ability level and by the characteristics of the analogy problems (vocabulary, relationship, and correspondence). The influence of these factors on performance is discussed separately for each of the three sections of the coding system (process analysis, strategy description, and evaluation).

Preliminary analyses were conducted to establish that the three groups were equal in ability as demonstrated by their GRE verbal scores, their scores on the paper-and-pencil analogy test, and their performance on the experimental analogies. As can be seen in Table 2, there were no differences among the groups on any of these measures. One-way analyses of variance on each of these measures produced F-values less than 1. Correlations among the three measures were significant as expected (see Table 2). The correlation between the GRE verbal subtest scores and the paper-and-pencil analogy test scores was higher than that between these scores and performance on the experimental analogies. This probably reflects, in part, differences in the methods of test administration, i.e., timed vs. untimed and paper-and-pencil vs. oral response.

## Process Description of Protocols

Dependent Variables. A detailed description of the dependent variables for this section of the coding system is presented in Table 3. First, six summary variables that represented the number of comments coded in each of the five process categories as well as the total number of comments coded were calculated for each subject for each item. Other summary variables, such as the number of options read aloud, had been included in the process coding system but were excluded from the category totals and treated separately. Finally, two additional variables, relationship development and option evaluation, were created.

These variables were examined to determine whether the independent variables of ability level and analogy characteristics had reliable effects on them. The discussion in the following sections focuses on those effects that were found to be relatively large and consistent across the three parallel sets of analogies.

The Effects of Ability Level on Performance. Subjects were classified as high or low ability on the basis of their GRE verbal scores (median = 555). In Figure 1 the mean frequency per analogy item for the five types of comments is presented as a function of ability level. The greatest number of the subjects' comments concerned the relationships between words. Comments about word meaning and decision processes also occurred with some frequency, while evidence of confirmation and executive processes was fairly infrequent. The primary effect of ability level was a tendency for low-ability subjects to comment more about word meaning than did high-ability subjects.

Table 2

Means and Standard Deviations for Three Measures of Verbal Skill for Three Groups and Correlations among Measures

Means and Standard Deviations

| | Group | | |
|---|---|---|---|
| Measure | 1 | 2 | 3 |
| GRE Verbal $\bar{X}$ | 537.5 | 542.5 | 542.5 |
| (SD) | (158.98) | (115.14) | (138.97) |
| Analogy Test $\bar{X}$ | 12.25 | 12.08 | 12.83 |
| (SD) | (3.47) | (4.52) | (4.28) |
| Experimental Analogies $\bar{X}$ | 4.75 | 5.42 | 4.50 |
| (SD) | (1.48) | (1.83) | (1.78) |

Correlations

| | GRE Verbal | Analogy Test |
|---|---|---|
| Analogy Test | .744** | |
| Experimental Analogies | .440* | .596** |

* $p < .04$

** $p < .0009$

Table 3

Dependent Variables in the Process Analysis

Category Summary Variables

- Encoding Comments
- Inference and Comparison Comments
- Decision and Response Selection Comments
- Confirmation Comments
- Executive Comments
- Total Number of Comments Scored

Other Summary Variables

- Relationship comparisons - number of options whose relationships were compared with the stem relationship
- Options read - number of options read aloud
- Options processed - number of options processed overtly (i.e., some comment was made about the options beyond repeating the words)
- Unexplained eliminations - number of options overtly eliminated without explicit reasons
- Justified eliminations - number of options overtly eliminated with stated reasons
- Number of words in protocol

Additional Variables

- Relationship development - sum of comments that were statements of the relationship between words in the stem or in an option or revisions of these relationships
- Option evaluation - sum of comments about whether or not an option was a probable correct answer, the number of options that were explicitly eliminated, and the number of options whose relationships were compared to the relationship of the stem

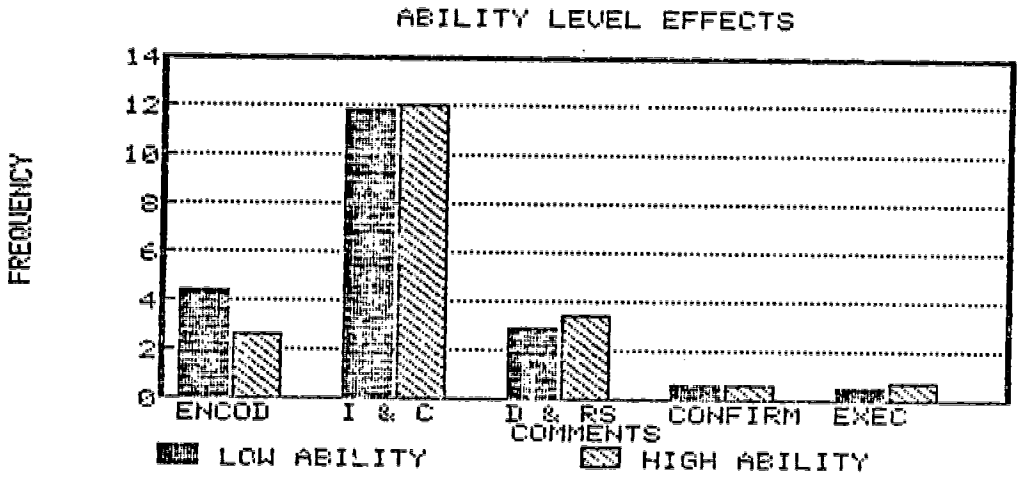



Figure 1. Mean frequency per analogy item of the five types of processing comments as a function of ability level. Encod = encoding, I & C = inference and comparison, D & RS = decision and response selection, Confirm = confirmation, Exec = executive.

The nine other summary variables were examined for large and consistent effects. The means and standard deviations for the most important effects are presented in Table 4. Ability level affected relationship comparisons and option evaluation. High-ability subjects compared the relationships for more option pairs with the stem pair than did low-ability subjects; they also made more comments indicative of option evaluation. There were tendencies for high-ability subjects to make more comments about the relationship between words in a pair, eliminate more options with explicit reasons, and, finally, to talk more when solving analogies.

The Effects of Analogy Characteristics on Performance. The analogy problems had varied factorially on three characteristics: vocabulary, relationship, and correspondence. The mean frequency per analogy item is presented for the five categories of comments in Figure 2 as a function of vocabulary (Figure 2a), relationship (Figure 2b), and correspondence (Figure 2c). Vocabulary had the largest effects. As can be seen in Figure 2a, more comments about encoding processes were made when analogies had difficult vocabulary than when they had easy vocabulary. In addition, more executive comments were made for the difficult vocabulary analogies. There was a tendency for more inference and comparison comments to be made for the easy vocabulary analogies. The primary effect of relationship was on executive comments, as more such comments were made for difficult relationship analogies than for easy relationship analogies. Finally, encoding comments were the only category affected by correspondence. More encoding comments were made for analogies with independent correspondence than for those with overlapping correspondence.

With respect to other summary variables, the number of options processed overtly was greater for analogies with easy vocabulary ($M = 4.26$) than for those with difficult vocabulary ($M = 3.77$). None of the other summary variables was affected by the analogy characteristics.

Summary of Process Analysis. Overall, comments indicative of inference and comparison processes were most frequent. Those indicative of encoding processes and of decision and response processes were more frequent than those about confirmation and executive processes. There was a trend for low-ability subjects to make relatively more encoding comments than did high-ability subjects. High-ability subjects compared option relationships with the stem relationship more frequently and made more comments evaluating options than did low-ability subjects. Thus, the picture that emerges is that low-ability subjects expend more effort in encoding analogy terms and less on comparing relationships and evaluating options than do high-ability subjects. A possible explanation for this pattern is that lower-ability subjects have poorer vocabularies. They may have more difficulty encoding terms and be less able to abstract relationships because of vocabulary difficulties.

Table 4

Means and Standard Deviations for Other

Summary Variables by Ability Level

| Variables | | Ability Level Low | High |
|---|---|---|---|
| Total Words | $\bar{X}$ | 145.05 | 190.97 |
| | SD | (60.21) | (99.22) |
| Relationship Comparisons | $\bar{X}$ | 2.13 | 3.15 |
| | SD | (1.86) | (1.00) |
| Justified Eliminations | $\bar{X}$ | .43 | .72 |
| | SD | (1.35) | (.58) |
| Relationship Development | $\bar{X}$ | 22.23 | 30.00 |
| | SD | (10.62) | (14.46) |
| Option Evaluation | $\bar{X}$ | 28.61 | 40.17 |
| | SD | (11.02) | (12.45) |

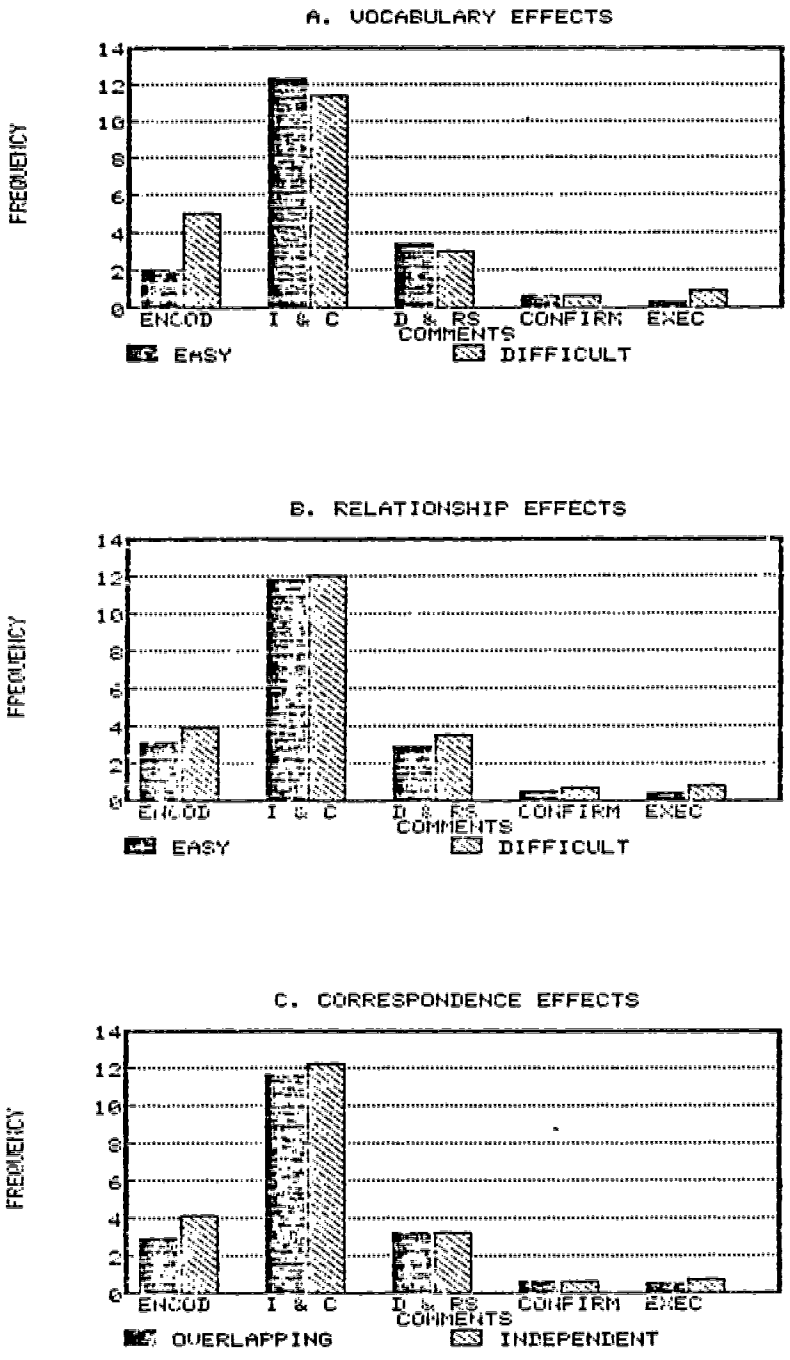



Figure 2. Mean frequency per analogy item of the five types of processing comments as a function of three analogy characteristics: A. vocabulary, B. relationship, and C. correspondence. Encod = Encoding, I & C = inference and comparison, D & RS = decision and response selection, Confirm = confirmation, Exec = executive.

The analogy characteristic which had the most impact on performance was vocabulary. Analogies with difficult vocabulary elicited relatively more encoding comments and relatively fewer comments about inference and comparison processes. In addition, more options were processed for analogies with easy vocabulary. This pattern parallels the one found to differentiate high and low ability subjects. Encoding processes are more evident when the analogy terms are difficult. Presumably, subjects do not expend much effort defining words with which they are familiar and when the vocabulary is easy, they can focus on inference and comparison processes and decision and response selection processes.

## Strategy Description

The second section of the coding system consisted of questions about the overall strategy subjects used to solve each analogy and more detailed questions about the approaches they used in specific situations (e.g., when they did not know the vocabulary). These questions are listed in Table 5 as is the percent of overall responses for each option. In addition, these data are presented separately for high- and low-ability groups and for the two dimensions of each of the three analogy characteristics (vocabulary, relationship, and correspondence). For those questions where one of the response options was "does not apply" the data reported for this category are the percents of instances for which the questions were not applicable. For the other options in such questions, the data are the percents of the total reponses to the remaining options. For example, in Question III, the data for options 1 through 4 represent the percent of responses for which the question applied while the data for option 5 is the percent of instances in which the question did not apply. For the purposes of discussion, only large differences (i.e., greater than 10 percent) between groups or levels of a factor will be considered. These differences are starred in the tables.

Question I concerned the overall strategy subjects used to solve each analogy. As can be seen in Table 5, the most commonly occurring strategy was the "ideal" one, or relationship centered. This strategy was more common for high-ability subjects than for low-ability ones, though even some of the latter used it more than 50 percent of the time. Vocabulary difficulty had a large effect on which strategy subjects used. The incidence of the use of the relationship-centered strategy was 25 percent higher for analogies with easy vocabulary than for those with difficult vocabulary, and there was a relatively high incidence of the use of unsystematic strategies in solving analogies with difficult vocabulary. Finally, correspondence also affected strategies. Relationship-centered strategies were more common for analogies with overlapping correspondence than for those with independent correspodence.

The second question concerned whether or not subjects used some kind of criteria to solve each analogy. In fact, subjects did so overwhelmingly

Table 5

Effects of Ability Level and Analogy Characteristics on Subjects' Strategies

| | Percent | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Analogy Characteristics | | | | | |
| | | Ability Level | | Vocabulary | | Rationale | | Correspondence | |
| | Overall | Low | High | Easy | Diff. | Easy | Diff. | Indep. | Overlap |
| I. Overall Strategy | | | | | | | | | |
| 1. Relationship Centered - Solution is characterized by a systematic approach from the outset: examinee formulates an Rs[1] at the outset or immediately identifies key and justifies with Rs or Ro= Rs[2], and then evaluates some/all options. | 62.2 | 54.9 | 69.4* | 75.0 | 49.3* | 66.7 | 57.6 | 54.2 | 70.1* |
| 2. Relationship Search - Solution is characterized by a relatively systematic approach from the outset: examinee attempts to formulate Rs at outset, states a vague or tentative Rs or states can't formulate Rs, then evaluates some/all options and develops Rs, Ro=Rs, or other stated criterion/criteria for solution. | 19.4 | 18.8 | 20.1 | 16.0 | 22.9 | 18.11 | 20.1 | 22.9 | 16.0 |
| 3. Alternative Criteria - Solution is characterized by a lack of evident systematic approach at the outset but during the course of exploring the item, examinee develops and applies some stated criterion/criteria for solution; examinee does not attempt to formulate Rs at the outset. | 10.8 | 15.3 | 6.3 | 8.3 | 13.2 | 9.0 | 12.5 | 11.8 | 9.7 |
| 4. Unsystematic - Solution is characterized by lack of evident systematic approach throughout: examinee does not formulate or does not attempt to formulate Rs at out set and does not develop and apply stated criterion/criteria during the course of exploring the item. | 7.6 | 11.1 | 4.2 | 0.7 | 14.6* | 6.3 | 9.0 | 11.1 | 4.2 |

[1]Rs - stem relationship

[2]comparison of option relationship (Ro) and stem relationship(Rs)

*difference between factor levels greater than or equal to 10%

Table 5 (cont.)

| | Percent | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Analogy Characteristics | | | | | |
| | | Ability Level | | Vocabulary | | Rationale | | Correspondence | |
| | Overall | Low | High | Easy | Diff. | Easy | Diff. | Indep. | Overlap |
| II. Application of criteria toward a solution (choose one) | | | | | | | | | |
| 1. appears to compare Ra and Ro, and/or to evaluate fit of R and words | 89.6 | 84.0 | 95.1* | 97.2 | 81.9* | 92.4 | 86.8 | 86.1 | 93.1 |
| 2. fails to apply any criteria in achieving solution | 10.4 | 16.0 | 4.9* | 2.8 | 18.1* | 7.6 | 13.2 | 13.9 | 6.9 |
| III. Discrimination among stated competing options (choose one) | | | | | | | | | |
| 1. evaluates competing options by appeal to already defined criteria (R, word fit, etc.) | 44.3 | 35.1 | 51.0* | 44.7 | 43.9 | 53.8 | 36.7* | 30.0 | 51.2* |
| 2. redefines criteria and reevaluates options | 30.7 | 32.4 | 29.4 | 31.9 | 29.2 | 33.3 | 28.6 | 40.0 | 25.6* |
| 3. chooses answer with articulated reason but not appeal to criteria | 9.1 | 8.1 | 9.8 | 10.6 | 7.3 | 2.6 | 14.3* | 0.8 | 11.6* |
| 4. arbitrarily chooses one option (reason not articulated) | 15.9 | 24.3 | 9.8* | 12.8 | 19.5 | 10.3 | 20.4* | 22.5 | 11.6* |
| 5. DNA | [69.4] | [74.3 | 64.6]* | [67.4 | 71.5] | [72.9 | 66.0] | [72.2 | 66.7] |
| IV. Evaluation and confirmation of final solution (choose one) | | | | | | | | | |
| 1. confirms solution | 54.5 | 58.3 | 50.7 | 51.4 | 57.6 | 50.0 | 59.0 | 55.6 | 53.5 |
| 2. does not confirm solution | 45.5 | 41.7 | 49.3 | 48.6 | 42.4 | 50.0 | 41.0 | 44.4 | 46.5 |

*difference between factor levels greater than or equal to 10%

Table 5 (cont.)

| | Percent | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Analogy Characteristics | | | | | | | |
| | | Ability Level | | Vocabulary | | Rationale | | Correspondence | |
| | Overall | Low | High | Easy | Diff. | Easy | Diff. | Indep. | Overlap |
| V. Strategies for compensating for apparent incomplete word knowledge (more than one may apply) | | | | | | | | | |
| 1. searches for meaning of unknown or partially known words | 30.4 | 27.4 | 34.1 | 27.9 | 30.9 | 32.1 | 28.9 | 32.2 | 27.3 |
| 2. infers meaning by appeal to other word(s) | 12.0 | 12.3 | 11.8 | 4.7 | 14.1* | 11.5 | 12.3 | 11.3 | 13.0 |
| 3. infers meaning by appeal to R | 14.1 | 11.3 | 17.6 | 16.3 | 13.4 | 15.4 | 13.2 | 13.9 | 14.3 |
| 4. analyzes word structure | 7.9 | 7.5 | 8.2 | 2.3 | 9.4 | 5.1 | 9.6 | 11.3 | 13.0 |
| 5. revises initial definition | 11.0 | 7.5 | 15.3 | 16.3 | 9.4 | 9.0 | 12.3 | 11.3 | 10.4 |
| 6. repeats word with no improvement in definition | 5.8 | 5.7 | 5.9 | 4.7 | 6.0 | 3.8 | 7.0 | 7.8 | 2.6 |
| 7. makes no attempt to figure out unknown words | 18.8 | 28.3 | 7.1* | 27.9 | 16.8* | 23.1 | 16.7 | 12.2 | 29.9* |
| 8. DNA | [58.0] | [49.3 | 66.7]* | [77.8 | 38.2]* | [61.8 | 54.2] | [52.8 | 63.2] |
| VI. Strategies for compensating for acknowledged incomplete derivation of Rs (choose one) | | | | | | | | | |
| 1. uses partially or tentatively defined Rs with no search to refine | 29.2 | 23.8 | 36.7* | 43.3 | 19.0* | 26.5 | 31.6 | 23.8 | 36.7* |
| 2. consciously explores options to derive or verify or refine Rs | 50.0 | 47.6 | 53.3 | 46.7 | 52.4 | 53.0 | 47.4 | 47.6 | 53.3 |
| 3. never develops Rs = solves item without R | 20.8 | 28.6 | 1.0* | 1.0 | 28.6* | 20.6 | 21.1 | 28.6 | 1.0* |
| 4. DNA | [75.0] | [70.8 | 79.2] | [79.2 | 70.8] | [76.4 | 73.6] | [70.8 | 79.2] |

*difference between factor levels greater than or equal to 10%

(in nearly 90 percent of the cases). Use of a criterion was affected by ability (high-ability subjects used one more frequently) and by vocabulary (use was more common for easier analogies).

The third question concerned the strategy subjects used to decide among two or more options that they considered possible correct answers. This situation occurred about 30 percent of the time. Under such circumstances subjects most frequently evaluated the options on the basis of previously defined criteria, although they also redefined the original criterion fairly frequently. The incidence of arbitrary selection of an option was 16 percent. Strategy selection in these circumstances was affected by ability, relationship, and correspondence. First, low-ability subjects stated that there were a number of potential correct answers less frequently than did high-ability subjects. Once they had identified a set of potential answers however, they tended to arbitrarily select one option more frequently and to evaluate the options on the basis of previously defined criteria less frequently than did high-ability subjects. Second, relationship difficulty also affected strategy. For analogies with difficult relationships, subjects tended to evaluate options by appealing to criteria less frequently and to arbitrarily choose an option or to choose an option for some reason other than a criterion reason more frequently. Finally, correspondence was found to affect the use of all four strategies. The incidence of appealing to previously defined criteria or other reasons was higher for analogies with overlapping correspondence. On the other hand, redefining criteria and arbitrarily choosing an option were more common for analogies with independent correspondence.

The strategies subjects used when they did not know the meaning of the vocabulary are described in Question V. This occurred about 42 percent of the time and was more common for low-ability subjects and for difficult vocabulary analogies. When subjects did not know a word, they most frequently made search-type comments (i.e., what does that word mean?) However, failure to make any attempt to figure out the meaning of the word was fairly common, especially among low-ability subjects and for analogies with easy vocabulary, as well as for analogies with overlapping correspondence. Two other common, strategies involved inferences based on other words or on relationships. Inferences based on other words were more common for difficult vocabulary items.

The final question in this section concerned the strategies subjects used when they recognized and stated that they did not completely understand the relationship between the stem words. This occurred in about 25 percent of the cases. When that did occur, subjects most frequently tried to infer the relationship from an exploration of the options. The second most common strategy was the application of a partial or incomplete stem relationship. This strategy was more common for high-ability subjects, for analogies with easy vocabulary, and for analogies with overlapping correspondence. The third strategy, solving the item without a relationship, was also fairly frequent and more common among low-ability subjects and for analogies with difficult vocabulary and with independent correspondence.

Summary of Strategy Description. This section of the coding system provided a qualitative description of the strategies that subjects used to solve the analogies and of more specific strategies that they used in certain conditions that constrained analogy solution. For the most part, subjects used a relationship-centered strategy that consisted of initially identifying the stem relationship and using it to guide or verify analogy solution. When subjects had difficulty in initially identifying the stem relationship, they often proceeded to examine the options and use this information to develop the stem relationship or some other criterion for solution. In only about 20 percent of the instances did subjects not attempt to formulate a stem relationship initially. In somewhat more than half these instances, subjects developed an alternative criterion for solution while they less frequently proceeded in an unsystematic and random manner. Thus, in nearly 90 percent of the instances, subjects adduced some criterion for their solution of the analogy.

Three situations were described that constrained analogy solution, and subjects' strategies in these conditions were categorized. These situations included identification of a set of more than one possible correct answer, ignorance of some of the vocabulary, and inability to understand the stem relationship clearly and/or completely. When subjects recognized a set of competing options, they typically selected the final answer by appealing to previously defined criteria or by redefining the criteria. Arbitrary selection of an option was less common. A variety of strategies were used by the subjects to compensate for incomplete word knowledge. Most commonly, subjects tried to recall the meaning of the word in question or to infer its meaning from other information in the analogy. However, failure to make any attempt to derive the word meaning was also fairly common. Finally, when subjects recognized that their formulation of the stem relationship was incomplete, they most frequently explored the options in order to better define the stem relationship. Another less frequently occurring strategy was the application of a partial relationship without further refinement. However, analogies were solved without the application of a stem relationship in about 21 percent of these instances.

Strategy was affected by ability level and by the characteristics of the analogies. As expected, high-ability subjects tended to use higher-level or more sophisticated strategies than did low-ability subjects. For example, when discriminating among competing options, high-ability subjects evaluated the options by appealing to a criterion more often than did low-ability subjects, while low-ability subjects arbitrarily choose an option more often than high-ability subjects did. As in the process analysis, the pattern for vocabulary tended to parallel that for ability group: higher-level strategies were more common for easy-vocabulary analogies. Relationship difficulty had surprisingly few effects on strategy. Although there was a tendency for relationship-centered strategies to be more common for easy-relationship analogies, this difference was not as striking as it was between the levels of the other variables. The only area where relationship difficulty had a strong impact was on how subjects discriminated among competing options. In this

situation, lower-level strategies were associated with difficult-relationship analogies. Finally, correspondence had effects that paralleled ability and vocabulary to a surprising degree. Here, overlapping correspondence was more frequently associated with higher-level strategies than was independent correspondence.

## Evaluation of Performance

The final section of the coding system consisted of the experimenters' judgments of the subjects' performance on the analogies in terms of the subjects' knowledge of the vocabulary and their expression of the stem relationship as well as the experimenters' evaluations of why a correct or incorrect answer was obtained. The results for this section are reported as percentages in Table 6. Again, the overall data are presented, and the data are also presented separately for ability groups and for analogy characteristics.

The first series of questions in this section concerned the experimenters' judgments of the subjects' knowledge of the vocabulary in the stem, the correct answer, and the other options. Subjects clearly understood the stem words in over 50 percent of the cases and apparently knew only one word in 19 percent of the cases. Knowledge of the stem words was related to all four of the independent variables. High-ability subjects appeared to know both stem words more frequently and know only one stem word less frequently than did low-ability subjects. The finding that subjects knew both stem words was much more frequent for easy vocabulary analogies than for difficult vocabulary items, while the opposite was true for the category "knows only one stem word." Similarly, knowledge of stem words was more common for analogies with easy relationships and with overlapping correspondence. In addition, subjects appeared to know only one stem word more frequently for analogies with independent correspondence. With respect to words in the correct answer, the experimenters were unable to score the subjects' knowledge in 19 percent of the instances. This effect was stronger for low-ability subjects and for analogies with difficult vocabularies, difficult relationships, and overlapping correspondence. On the other hand, in a high percent of the scorable instances (72.5 percent), subjects appeared to know both key words. This effect was greater in high-ability subjects, and for easy-vocabulary analogies and easy-relationship analogies. The final question in this section concerned knowledge of words in the remaining options. This information was unscorable in nearly 30 percent of the instances, while subjects appeared to know all the words in more than 55 percent of the instances. Once again, the incidence of unscorability was higher for low-ability subjects and for difficult-relationship analogies. High-ability subjects appeared to know all the words more frequently than did low-ability subjects. Words were known less frequently for difficult-vocabulary analogies and for independent correspondence analogies. It is important to note that

Table 6

Evaluation of Performance

| | Percent | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Analogy Characteristics | | | | | |
| | | Ability | | Vocabulary | | Relationship | | Correspondence | |
| | Overall | Low | High | Easy | Diff. | Easy | Diff. | Indep. | Overlap |
| I. Word Knowledge | | | | | | | | | |
| A. Stem | | | | | | | | | |
| 1. appears to know both stem words | 56.5 | 44.6 | 71.3* | 81.0 | 44.5* | 72.1 | 52.9* | 54.0 | 70.4* |
| 2. appears to know only one stem word | 19.0 | 24.7 | 12.0* | 7.5 | 32.9* | 17.7 | 23.2 | 25.3 | 15.8* |
| 3. doesn't know either stem word | 3.6 | 4.3 | 2.7 | 0.0 | 7.7 | 2.0 | 5.8 | 6.0 | 2.0 |
| 4. knowledge of stem word(s) vague | 6.5 | 7.0 | 6.0 | 4.8 | 7.1 | 4.1 | 7.7 | 6.7 | 5.3 |
| 5. knowledge of stem word(s) clearly inaccurate | 3.3 | 4.3 | 2.0 | 2.0 | 2.6 | 1.4 | 3.2 | 2.7 | 2.0 |
| 6. uses inappropriate but valid meaning for stem word(s) | 3.3 | 5.9 | 0.0 | 2.7 | 0.6 | 0.7 | 2.6 | 1.3 | 2.0 |
| 7. infers correct meaning of stem word(s) after consideration of other information | 4.5 | 4.3 | 4.7 | 1.4 | 3.9 | 1.4 | 3.9 | 4.0 | 1.3 |
| 8. knowledge of stem words unscorable | 3.3 | 4.8 | 1.3 | 0.7 | 0.6 | 0.7 | 0.6 | 0.0 | 1.3 |
| B. Actual Key | | | | | | | | | |
| 1. appears to know both words in key | 72.5 | 60.8 | 84.0* | 80.4 | 64.6* | 83.9 | 61.0* | 69.4 | 75.5 |
| 2. appears to know only one word in key | 3.8 | 6.3 | 1.4 | 1.4 | 5.6 | 0.7 | 6.3 | 4.9 | 2.1 |
| 3. doesn't know either key word | 0.3 | 0.7 | 0.0 | 0.0 | 0.7 | 0.0 | 0.7 | 0.7 | 0.0 |
| 4. knowledge of key word(s) vague | 2.8 | 3.5 | 2.1 | 4.2 | 1.4 | 2.1 | 3.5 | 2.1 | 3.5 |

*difference between factor levels greater than or equal to 10%

Table 6 (cont.)

| | Percent | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Analogy Characteristics | | | | | |
| | | Ability | | Vocabulary | | Relationship | | Correspondence | |
| | Overall | Low | High | Easy | Diff. | Easy | Diff. | Indep. | Overlap |
| I. Word Knowledge (cont.) | | | | | | | | | |
| B. Actual Key (cont.) | | | | | | | | | |
| 5. knowledge of key word(s) clearly inaccurate | 1.0 | 0.0 | 2.1 | 1.4 | 1.4 | 1.4 | 1.4 | 0.0 | 2.8 |
| 6. uses inappropriate but valid meaning for key word(s) | 0.7 | 0.7 | 0.7 | 0.0 | 1.4 | 0.0 | 1.4 | 0.0 | 1.4 |
| 7. infers correct meaning of key word(s) after consideration of other information | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 8. knowledge of key words unscorable | 18.8 | 28.0 | 9.7* | 12.6 | 25.0* | 11.9 | 25.7* | 2.3 | 14.7* |
| C. Other options | | | | | | | | | |
| 1. appears to know all words in other options | 55.5 | 47.5 | 63.8* | 66.2 | 44.7* | 57.4 | 53.5 | 50.7 | 60.3* |
| 2. doesn't know some words in other options | 14.8 | 17.0 | 12.8 | 8.5 | 22.3* | 17.7 | 12.0 | 16.2 | 13.5 |
| 3. doesn't know any words in other options | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4. knowledge of words in other options unscorable | 29.7 | 36.2 | 23.5* | 25.4 | 34.0 | 24.8 | 34.5* | 33.1 | 26.2 |

*difference between factor levels greater than or equal to 10%

Table 6 (cont.)

| | Percent | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Analogy Characteristics | | | | | |
| | | Ability | | Vocabulary | | Relationship | | Correspondence | |
| | Overall | Low | High | Easy | Diff. | Easy | Diff. | Indep. | Overlap |
| II. Degree of correct final formulation of Stem Relationship (R) | | | | | | | | | |
| 1. formulates R correctly at outset | 47.9 | 39.6 | 56.3* | 55.9 | 39.9* | 60.6 | 35.4* | 42.7 | 53.1* |
| 2. formulates R correctly after consideration of option pairs | 10.8 | 8.3 | 13.4 | 11.2 | 10.5 | 7.7 | 13.9 | 10.5 | 11.2 |
| 3. formulates R partially correctly at outset | 10.5 | 10.4 | 10.6 | 13.3 | 7.7 | 4.9 | 16.0* | 11.9 | 9.1 |
| 4. formulates R partially correctly after consideration of option | 5.2 | 4.9 | 5.6 | 2.8 | 7.7 | 2.8 | 7.6 | 3.5 | 7.0 |
| 5. formulates R incorrectly at outset | 4.2 | 6.9 | 1.4 | 4.2 | 4.2 | 1.4 | 6.9 | 3.5 | 4.9 |
| 6. formulates R incorrectly after consideration of option pairs | 9.8 | 11.1 | 8.5 | 8.4 | 11.2 | 11.3 | 8.3 | 12.6 | 7.0 |
| 7. Never formulates R | 11.5 | 18.8 | 4.2* | 4.2 | 18.9* | 11.3 | 11.8 | 15.4 | 7.7 |
| III. Correct answer (choose one) | | | | | | | | | |
| 1. correct relationship and vocabulary | 73.1 | 63.5 | 80.2* | 82.1 | 62.5* | 75.7 | 69.1 | 67.1 | 78.1* |
| 2. correct relationship, weak vocabulary | 9.7 | 10.8 | 8.9 | 4.2 | 16.3* | 8.4 | 11.8 | 11.4 | 8.3 |
| 3. weak relationship, correct vocabulary | 4.6 | 4.1 | 5.0 | 7.4 | 1.3 | 4.7 | 4.4 | 5.1 | 4.2 |
| 4. weak relationship and vocabulary | 10.9 | 20.3 | 4.0* | 6.3 | 16.3* | 9.3 | 13.2 | 12.7 | 9.4 |
| 5. chance guess - (uninformed, no processing of item) | 1.7 | 1.4 | 2.0 | 0.0 | 3.8 | 1.9 | 1.5 | 3.8 | 0.0 |
| 6. DNA | 38.8 | 48.6 | 28.9* | 33.6 | 44.1* | 24.6 | 52.8* | 44.8 | 32.9* |

*difference between factor levels greater than or equal to 10%

Table 6 (cont.)

| | Percent | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Analogy Characteristics | | | | | |
| | | Ability | | Vocabulary | | Relationship | | Correspondence | |
| | Overall | Low | High | Easy | Diff. | Easy | Diff. | Indep. | Overlap |
| IV. Incorrect answer (heuristic) | | | | | | | | | |
| 1. incorrect formulation of stem or option relationship Rs or Ro given mastery of vocabulary | 22.8 | 16.9 | 32.6* | 35.4 | 13.6: | 19.4 | 24.4 | 22.4 | 23.4 |
| 2. incorrect application of reasonable stem or option relationship Rs or Ro | 25.4 | 26.8 | 23.3 | 35.4 | 18.2* | 25.0 | 25.6 | 25.4 | 25.5 |
| 3. ignorance of vocabulary in stem and/or options prevents development of relationship | 20.2 | 22.5 | 16.3 | 12.5 | 25.8* | 22.2 | 19.2 | 22.4 | 17.0 |
| 4. ignorance of vocabulary leads to incorrect relationship | 28.1 | 28.2 | 27.9 | 16.7 | 36.4* | 30.6 | 26.9 | 25.4 | 31.9 |
| 5. chance guess (uninformed, no processing of item) | 3.5 | 5.6 | 0.0 | 0.0 | 6.1 | 2.8 | 3.8 | 4.5 | 2.1 |
| 6. DNA | 60.3 | 50.7 | 69.4* | 66.4 | 54.2* | 74.8 | 45.8* | 53.5 | 67.1* |

*difference between factor levels greater than or equal to 10%

the high incidence of unscorability for key and option words constrains the interpretation of these results. For example, it is possible (but not probable) that low-ability subjects knew as many key words as did high-ability subjects; however, they did not provide any overt evidence of this knowledge in their protocols.

The experimenters' judgments of how well subjects construed the stem relationship is reported in Question II. Overall, subjects construed this relationship correctly either initially or after consideration of the options about 59 percent of the time. On the other hand, subjects did not formulate the stem relationship at all 11 percent of the time. The correct relationship was formulated more frequently by high-ability subjects than low-ability subjects. In addition, low-ability subjects were unable to formulate the relationship more frequently than were high-ability ones. Similar effects were found for vocabulary. Relationships were formulated correctly more frequently for easy-vocabulary analogies than for difficult-vocabulary ones, and there was a higher incidence of failure to formulate any relationship for the difficult-vocabulary analogies. Relationship also affected performance. As might be expected, the relationship was formulated correctly more often when it was easy than when it was difficult. Furthermore, the development of relationships that were only partially correct was more common when the relationships were difficult. Finally, more initially correct relationships were formulated for analogies with overlapping correspondence than for independent ones.

Question III concerned the experimenters' evaluations of how subjects arrived at correct responses. Overall, subjects were correct slightly more than 60 percent of the time. Correct answers were seldom the result of random guessing. When subjects were correct, they appeared both to know the vocabulary and to construe the relationship correctly more than 70 percent of the time. High-ability subjects were correct more often than low-ability subjects. However, it is interesting to note that there were differences in how the two groups arrived at correct answers. Low-ability subjects had a lower incidence of analogies solved correctly on the basis of both good knowledge of the vocabulary and good statements of the relationships, and they had a higher incidence of correct solutions on the basis of weak vocabulary and relationships. Vocabulary also had large effects on how analogies were solved. First, about 10 percent more analogies with easy vocabulary were solved correctly than those with difficult vocabulary. Given a correct solution, difficult-vocabulary analogies, when compared with easy-vocabulary analogies, were solved less frequently on the basis of both correct relationships and good word knowledge and more frequently on the basis of correct relationships despite weak vocabulary and of both weak relationships and vocabulary. Relationship difficulty had the largest effect of any factor on the correctness of analogy solution in that more than 50 percent of the analogies with difficult relationships were not solved correctly.

However, relationship difficulty did not affect the basis for correct solutions. Finally, analogies with overlapping correspondence were solved correctly more often, and were solved correctly on the basis of both correct relationships and vocabulary more often, than were analogies with independent correspondence.

The last question in this section concerned the basis for incorrect solutions. The frequency of incorrect solutions is the complement of the frequency of correct solutions described above and will not be restated here.

Overall, incorrect solutions were due with nearly equal frequency to incorrect formulations of relationships when vocabulary was known, incorrect application of well-formulated relationships, failure to formulate relationships, or incorrect formulation of relationships due to ignorance of vocabulary. Wrong answers were seldom the result of random guessing. The most frequently occurring source of errors for high-ability subjects was an incorrect relationship given mastery of the vocabulary, while low-ability subjects most often generated incorrect relationships because of vocabulary difficulties or incorrectly applied reasonable relationships. Vocabulary difficulty had multiple influences on the basis for incorrect solution. For easy-vocabulary analogies, incorrect solutions were most frequently the result of incorrect formulations of relationships or the incorrect application of reasonable relationships. On the other hand, incorrect solutions of difficult-vocabulary analogies were based on ignorance of the vocabulary preventing the development of or leading to incorrect formulations of the relationships.

Summary of Evaluation of Performance. Questions in this section of the scoring system focused on the subjects' apparent knowledge of the vocabulary in the analogy items, their mastery of the central relationships in the items, and the factors contributing to their selection of correct or incorrect answers.

With regard to word knowledge, subjects demonstrated some difficulty with stem vocabulary in about 40 percent of the instances but with key vocabulary only about 8 percent of the time. As expected, low-ability subjects had more difficulty with vocabulary than did high-ability subjects, and all subjects had more problems with the words in difficult-vocabulary analogies than in the easy-vocabulary ones. However, the finding that relationship and correspondence were also related to word knowledge was unexpected and suggests that the experimenters' attempts to counterbalance vocabulary difficulty with the other two stimulus characteristics were not entirely successful.

Overall, subjects formulated a correct or partially correct relationship initially in about 58 percent of the instances and, after consideration of the options, in about 16 percent of the instances; they were unable to formulate the relationship or formulated it incorrectly about 26 percent of the time. Once again, ability level and vocabulary

had similar effects. As might be expected, relationship difficulty also influenced the formulation of the stem relationship, as did correspondence.

All the independent variables affected the frequency of correct solution. High-ability subjects, easy vocabulary, easy relationships, and overlapping correspondence were all associated with a higher success rate. Correct answers were most frequently based on good knowledge of the vocabulary and well-formulated relationships. However, low-ability subjects had a relatively high rate of correct solutions based on weak vocabulary and poorly formulated relationships. For difficult-vocabulary analogies the incidence of correct solutions despite poor word knowledge was relatively high. With respect to the basis for incorrect responses, the most interesting findings concerned vocabulary. Difficult vocabulary led to an incorrect formulation of the relationship relatively frequently. Although relationship difficulty had the largest impact on probability of an incorrect solution, it did not differentially affect the way that such solutions were attained.

## Discussion

### Relationship of Findings to Previous Research

The results of this study generally are consistent with findings of other studies of cognitive processes in the solution of verbal analogy problems. Previous studies employing protocol analysis methods (Conolly & Wantman, 1964; Heller, 1979; Heller & Pellegrino, 1978; Whitely & Barnes, 1979) have found that subjects' organization of problem solving is an important factor in the successful solution of verbal analogies. Cognitive component studies (Sternberg, 1977, for example) have identified a small number of information processes that underlie solution of analogy problems. The present study drew on these analyses of problem solving. Oral comments made by subjects during problem solving were analyzed in terms of the cognitive activities highlighted in cognitive components research, albeit with modifications to fit the nature of the protocol data and problems under analysis. In addition, subjects' problem-solving activities were characterized in terms of global strategies that have also been investigated in previous studies.

Consistent with earlier findings in cognitive research, we found that subjects who were good at solving analogy problems tended to make more inferences concerning relationships and to compare stem and option relationships more. In contrast to previous research (e.g., Sternberg, 1977), we found evidence that high-ability subjects concentrate less on the encoding of the problem information than do low-ability subjects. This apparent disagreement with previous research may be due to methodological differences between the present study and chronometric studies

as well as differences in the concept of "encoding" across studies. There is some ambiguity about whether the concept of encoding applies to the encoding of the meaning of stem words alone, or includes encoding the words in the options as well as formulation of the stem relationship. In this study, encoding comments were restricted to overt, verbalized evidence that subjects were actively processing or searching for the meanings of either stem or option words. We did not record encoding as having occurred in the absence of such overt evidence even if it had obviously occurred covertly. One important methodological difference between this study and most other studies of analogies was in the presentation format of the analogy problems. In most other studies, the analogies were of the form A:B::C: ____, while in the present study only the A:B pair was given and the C:D pair, rather than just the D term, had to be selected from among the options. In addition, it is also important to note that the dependent variables differ between this study (frequency of overt comments) and chronometric studies (time to execute a hypothetical process). Thus, the relationship between our results and those of chronometric studies is indirect, though a reasonable degree of agreement can be expected.

Encoding comments in our study occurred more frequently when analogies involved difficult vocabulary. We found that subjects were more thorough in their processing of options when vocabulary was easy.

Consistent with previous research (Heller & Pellegrino, 1978; Rumelhart & Abrahamson, 1973,) we found that subjects were influenced by the occurrence of semantic connections in the meanings of words making up analogy problems. We found that analogies with overlapping semantic correspondences among stem and key terms were more likely to be solved successfully on the basis of correctly inferred relationships and vocabulary. This finding is consistent with the previous research cited, which found that subjects judge completed analogies as more appropriate when their terms are semantically related, and that they are more likely to solve analogies involving semantically related terms.

Our results are consistent with previous findings suggesting that high-ability individuals are more organized in their analogy problem solving than are those with low ability. Like Heller (1979) and Alderton, Goldman, and Pellegrino (1985), we found that high-ability subjects tended to be more systematic and comprehensive in their analysis of problem information and in their decision making.

## Implications for Test Development

The results of the present study have several implications for test development. While test developers may have intuitively anticipated a number of the findings, concrete evidence of certain features of analogy items permits them to turn intuitions of varying degrees of plausibility into guidelines for item and test development. In addition, the study provides much information that could not have been obtained without investigating actual examinee performance. Finally, the protocols provide a unique opportunity for test developers to witness test takers'

minds at work; this can be of great human as well as practical interest. The implications of our results for writing instructional materials for examinees and for test and item development are discussed in the following sections.

Instructional Materials. The results would suggest that most examinees understand the basic task of analogy items and, on the whole, approach the solution of analogies sensibly. The fact that 62 percent of the time subjects utilized an "ideal" strategy (first deriving and then applying the stem relationship) and that 82 percent of the time their strategies involved the derivation and application of a word-pair relationship would suggest that they understood the general requirements of analogy solution. Furthermore, the process analysis suggests that high- and low-ability examinees do not differ markedly (with the exception of encoding) in the basic activities in which they engage to solve analogies. This has relevance to instructional or self-help materials prepared for examinees (such as the Information Bulletin): such materials would be improved if they helped examinees to refine this basic strategy by suggesting that they develop a set of secondary strategies. (Of course, such materials should continue to emphasize the importance of systematic solution of analogies based upon evaluation of word-pair relationships.)

In addition to emphasizing the overall strategy, the brief "tips" section in the Information Bulletin includes two other suggestions that the results described above suggest are associated with success on GRE analogies: considering each option and reevaluating the stem relationship when more than one option appears plausible. These suggestions could be expanded and strengthened on the basis of study results to include more emphasis on (1) evaluating the fit of the stem relationship to each of the option pairs (rather than simply "reading" each answer choice, as the Bulletin suggests), wherever possible arriving at a decision with regard to the appropriateness of that option (e.g., definitely not the answer, possibly the answer, etc.), and wherever possible articulating a reason for that decision, and (2) the usefulness of first eliminating the most implausible options and then refining the definition of the stem relationship to help discriminate among the remaining options.

Other "tips" or strategies were suggested in part by the findings discussed above and in part by notes made by the experimenters as they scored protocols. Examinees should be encouraged to define the stem relationship in as precise detail as possible; for example, for the stem pair FLOWCHART:PROCEDURE, subjects who defined the relationship "a flowchart shows you a procedure" were more likely to be tripped up by a close distracter in the item than were subjects who defined the relationship more precisely, for example, "a flowchart is a graphic representation of a procedure." Examinees should be instructed to pay attention to the part of speech of the words in the stem pair when formulating the relationship between the words; trying to formulate a relationship using one (or both) of the words in a part of speech different from that of the

given word is likely to create obstacles to the correct application of the relationship. A number of other suggestions for examinees might be derived from the protocols with further analysis. The usefulness of making associations between the first words or the second words in the stem and an option (rather than associating the relationships) might be discussed--it is, however, a strategy to be treated with great care. Another area for possible discussion would be strategies for working with items in which some words are unfamiliar or the relationships difficult to derive.

Test and Item Development. Perhaps the most salient features of these results for test and item development are the considerable effect of relationship difficulty on whether an item is solved correctly and the considerable effect of vocabulary difficulty on how an item is solved. These findings relate not only to the development of items but to general considerations about what analogies appear to measure and how the analogy item type might be modified.

The finding that relationship difficulty appears to have considerable effect on whether an item is solved correctly--that is, on the difficulty of an item--is reassuring to the test development concern that analogies not be primarily a "vocabulary" item type. However, the fact that the effect of relationship difficulty was confounded to a greater degree than we had expected with vocabulary difficulty prevents us at this point from drawing further conclusions about the role of relationship difficulty on item difficulty. However, one interesting finding of the study is that relationship difficulty has considerably less impact than vocabulary difficulty on the way in which examinees approach analogy solution.

Of the three stimulus dimensions, vocabulary had the greatest number of effects on the activities described in the process analysis as well as on the strategies employed by subjects. Poor vocabulary appears to have been the major factor in about half the instances of incorrect solutions to items. In addition, difficult vocabulary seemed to "derail" subjects more than did difficult relationships. When vocabulary was difficult, subjects made fewer inference and comparison comments, processed fewer options, and employed the "ideal" strategy less often. Difficult relationships did not have this effect of lowering the incidence of "desirable" behavior, nor did easy relationships have the opposite effect of encouraging these behaviors, as did easy vocabulary items. When subjects failed to formulate any relationships at all in the solution of items, vocabulary had a large effect and relationship almost none at all.

Such a pattern could reasonably have been expected. When examinees do not know a word in a word pair, they are unlikely to make up an arbitrary relationship and apply it to other word pairs; rather, they will engage in other activities to compensate for their vocabulary difficulties. Examinees generally know when they do not know the meaning of a word. They less frequently recognize when they have not fully grasped the

relationship between words in a pair. Our findings suggest that difficult relationships seem less likely than difficult vocabulary to distract examinees from what they appear to recognize as good analogy-solving behavior; it might be said that the subjects in this study were less intimidated by difficult relationships than by difficult vocabulary.

Theoretically, if one wished to emphasize the reasoning component in analogy items, one would eliminate difficult words and try to compensate for the loss in difficulty contributed by vocabulary by, for example, increasing the number of items with difficult relationships and reducing the number of overlapping items. Practically speaking, it would be a challenging but not impossible task to produce the requisite numbers of high-difficulty items for the GRE Program while reducing the use of difficult words. It would be challenging in part because difficult relationships often depend upon precision or fine distinctions in definition. In addition, difficult items depend upon the use of close distracters, which also often depend upon fine distinctions in definition. The vocabulary issue might thus reappear in different guise. In other words, vocabulary and relationship difficulty may be inextricably linked in analogy items that are appropriate for the GRE. Further research that more rigorously analyzed the impact of vocabulary and relationship difficulty on item difficulty would help to illuminate this issue. Such research would permit more informed discussion about what the analogy item type should measure, whether analogy items currently being developed are appropriate, and whether modifications are feasible.

However, the findings of this study with respect to the roles of relationship and vocabulary suggest one modification in analogy item writing that is possible at present: reducing the incidence of difficult vocabulary words in the <u>stems</u>. Since examinees appear to be more easily "derailed" by difficult vocabulary than by difficult relationships, this modification would probably lead to greater examinee concentration on developing and applying the stem relationship. One possible conclusion from the study data is that examinees are less bothered by words they do not know in options than they are by words they do not know in the stem—quite reasonably, since knowledge of stem words is essential to developing a reliable relationship on the basis of which the item is solved. This modification would give greater emphasis to the reasoning component in analogy items, and would probably increase examinee perceptions of analogies as more a measure of verbal <u>reasoning</u> than of vocabulary. However, it should be borne in mind that it might prove difficult to produce the requisite number of high difficulty items if such a modification were made.

The relationship of correspondence to subject performance seems in general to parallel that of vocabulary, with overlapping items producing some of the same effects that easy vocabulary items did. Again, it might reasonably have been expected that overlapping items would be associated with greater efficiency in solving analogies since the presence of

overlap often provides the examinee with a basis other than relationship on which to make associations between stem and correct answer. The findings relating to correspondence are useful for test development staff, who will now have a better understanding of its role.

With regard to current test specifications, our study tends to support the usefulness of the correspondence category. On the other hand, it also suggests that ratings of vocabulary and relationship difficulty might be appropriate elements to consider in classifying items. It would probably be no more difficult to achieve consensus on what constitutes, for example, an easy or difficult relationship than it currently is for item writers to achieve consensus on what constitutes an independent or overlapping item. On the other hand, the very process of assembling a GRE test--which involves choosing items of a range of difficulty--almost inevitably results in a range of vocabulary and relationship difficulty among the items since these features appear to be major contributors to item difficulty.

In terms of item development, the results of this study confirm the general strategies utilized by test developers and suggest possible refinements. In a sense, for every "tip" or suggestion that is provided to instruct examinees, there is a corresponding strategy item writers can employ. Most obviously--with regard to the stimulus dimensions investigated in this study--item writers can, to a certain extent, control vocabulary level, complexity of relationship, and correspondence, as well as combinations of these dimensions. At a more refined level, they can control the number of close distracters, thereby eliciting more or less refining of the stem relationship by examinees: two or more options can be developed that parallel a general relationship borne by the stem words, while only one of the options parallels a more detailed statement of the stem relationship. Item writers can manipulate the semantic relationship between the two stem terms: for example, the more natural part of speech in the definition of a word-pair relationship might be changed to add complexity to the relationship, or the more natural order of words in a relationship might be reversed. Item writers can also, with care, develop distracters that work off of associations among first words (or among second words) across two or more pairs whose relationships do not match. As with advice to examinees, other suggestions for item writers might be elicited from further analysis of the protocols.

Because of the pilot nature of this study, only certain characteristics of analogy items were examined in relation to examinee performance. The study did not undertake to examine in a rigorous way the interrelationships among these characteristics. Other features of analogies might be investigated--for example, areas of the content specifications that were not incorporated in this study and dimensions of word-pair and interword relations. The protocols collected for this study themselves could be analyzed for further information, such as the relationship of examinee perceptions of difficulty to actual difficulty and to item characteristics, or the characteristics of wrong answer choices.

The Usefulness of Protocol Methodology

Collection and analysis of verbal comments made by subjects in this study proved productive, and the results of the study suggests that similar procedures could be used in analyzing thinking skills used by examinees solving other GRE verbal item types. One major benefit derived from the use of protocol analysis stands out. Use of protocol methods permitted systematic collection of evidence on what examinees think about as they solve GRE analogy items. The protocols we have collected are a valuable resource for the GRE Program, GRE test development staff, and cognitive researchers. The data, for the first time, document examinees' perceptions of analogy items and the problem-solving processes that underlie solution of analogies.

Of course, one can question whether the collected protocols legitimately represent what examinees actually do when confronted with analogy problems under actual testing conditions. There simply is no way to answer this concern directly. Any scientific method of performance analysis requires an assessment intervention, and this might entail a distortion of the phenomena under study. We believe, however, that our findings do, in large part, represent important aspects of what examinees actually do as they solve analogy problems in the GRE. Other research might be conducted to investigate whether variations in protocol analysis methods lead to evidence of behavior similar to or different from what was found in the present study.

To the extent our protocol methods do not represent a serious distortion of cognitive behaviors underlying solution of GRE analogy items, we have helped in evaluating the construct validity of the analogy item type. Our findings make sense in light of other cognitive research on verbal analogy problem solving, and thus we can begin to have some confidence that we know what analogy items measures based on cognitive research and theory.

Our approach to protocol analysis was in large part successful because we drew on previous cognitive research to aid us in developing our analytical approach, though it must be emphasized that our knowledge of test development and familiarity with the GRE analogy item type contributed as well to our approach and analysis. This combination of cognitive theory and test development approaches would appear important to future efforts to analyze performance on other GRE verbal item types.

A final comment should be made about future directions for research in this area. Construct validity studies of the sort described here are valuable in that they document the basic cognitive processes that underlie performance on items in college and graduate school admissions tests. A further challenge exists, however, in more completely establishing the

construct validity of such items. This challenge is to demonstrate that there are direct and verifiable connections between the cognitive processes required to solve test items and the cognitive processes required to perform actual academic tasks in college and graduate school. Thus, links between psychological theories of cognitive processes, admissions testing, and instructional practice might be developed. Such an integration of theory, testing, and instruction would support the development of testing instruments that would be an integral part of instructional practice and special programs designed to systematically develop in students those cognitive skills necessary for higher education.

References

Alderton, D. L., Goldman, S. R., & Pellegrino, J. W. (1985). Individual differences in process-outcomes for verbal analogy and classification solution. Intelligence, 9, 1-14.

Connolly, J. A., & Wantman, M. J. (1964). An exploration of oral reasoning processes in responding to objective test items. Journal of Educational Measurement, 1, 59-64.

Ericson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Boston: MIT Press.

Glaser, R., & Pellegrino, J. W. (1978). Uniting cognitive process theory and differential psychology: Back home from the wars. Intelligence, 2, 305-319.

Goldman, S. R., & Pellegrino, J. W. (1984). Deductions about induction: Analyses of developmental and individual differences. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence, (Vol. 2). Hillsdale, NJ: Erlbaum.

Graduate Record Examinations Board (1983). GRE Information Bulletin. Princeton, NJ: Educational Testing Service.

Heller, J. I. (1979). Cognitive processing in verbal analogy solution. Unpublished doctoral dissertation, University of Pittsburgh, PA.

Heller, J. I., & Pellegrino, J. W. (1978, March). Cognitive processes and sources of item difficulty in the solution of verbal analogies. Paper presented at the meeting of the American Educational Research Association, Toronto.

Hunt, E. (1978). Mechanics of verbal ability. Psychological Review, 85, 109-130.

Newell, A., & Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice Hall.

Pellegrino, J. W., & Glaser, R. (1980). Components of inductive reasoning. In R. E. Snow, P. A. Federico, & W. E. Montague, (Eds.), Aptitude, learning, and instruction: Cognitive process analyses of aptitude (Vol. 1). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed.), Advances in instructional psychology. Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. Cognitive Psychology, 5, 1-28.

Sternberg, R. J. (1977). Intelligence, information processing, and analogical reasoning. New York: Halsted Press.

Sternberg, R. J. (1982). Reasoning, problem solving, and intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence. New York: Cambridge University Press.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. GREB No. 79-8P. Princeton, NJ: Educational Testing Service.

Whitely, S. E. (1980). Modeling aptitude test validity from cognitive components. Journal of Educational Psychology, 72, 750-769.

Whitely, S. E., & Barnes, G. M. (1979). The implications of processing event sequences for theories of analogical reasoning. Memory & Cognition, 7, 323-331.

Wilson, K. (1985). The relationship of GRE General Test item-type part scores to undergraduate grades. GREB No. 81-22P/ETS RR No. 84-38. Princeton, NJ: Educational Testing Service.